FACULTY OF DATA AND DECISION SCIENCES
MAKING SENSE OF IT

הטכניון מכון טכנולוגי לישראל TECHNION Israel Institute of Technology

Classifying Jobs Listings as Real or Fake Using Semi-Supervised Machine Learning

Shachaf Haviv, Shani Angel, Nitzan Manor, Saar Manshrov

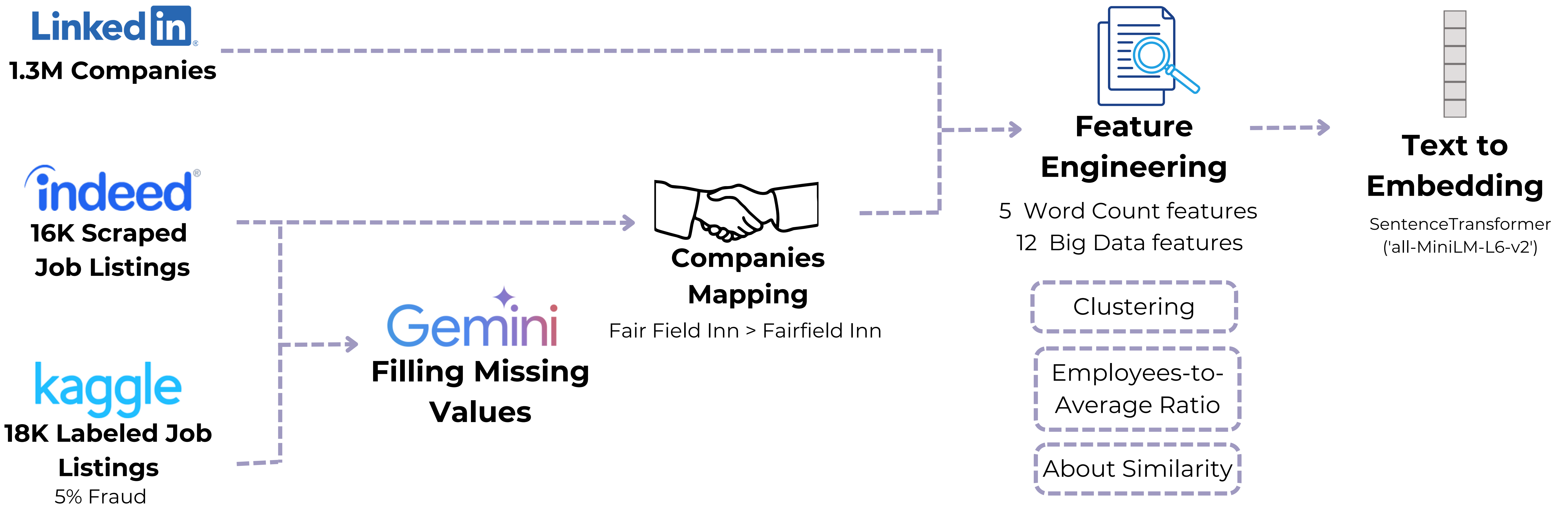
Motivation

FAKE **36% of job listings are fake** REAL
according to Forbes

Goal

Utilize Big data and ML to detect fake job listings

Data Process



1.3M Companies

16K Scraped Job Listings

18K Labeled Job Listings
5% Fraud

Filling Missing Values

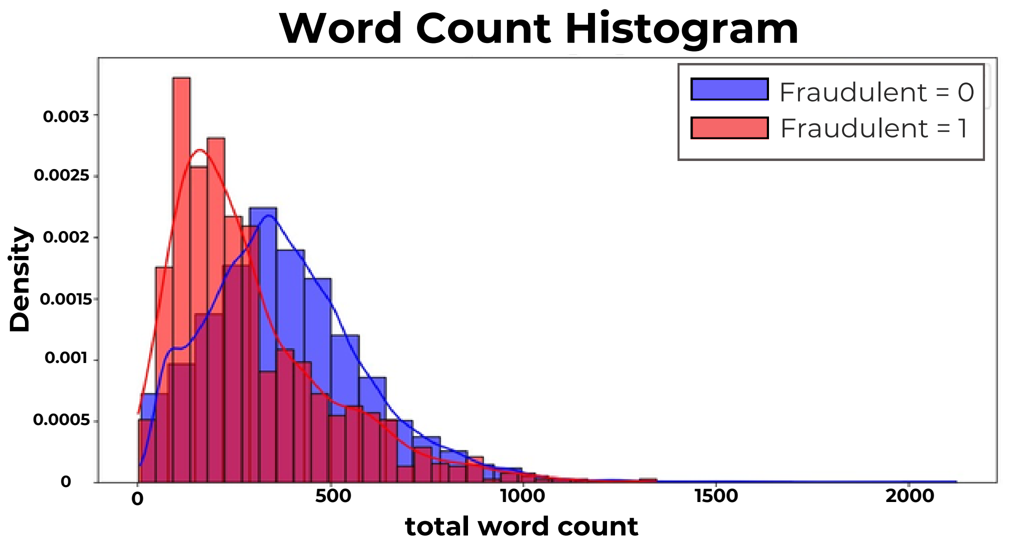
Companies Mapping
Fair Field Inn > Fairfield Inn

Feature Engineering
5 Word Count features
12 Big Data features

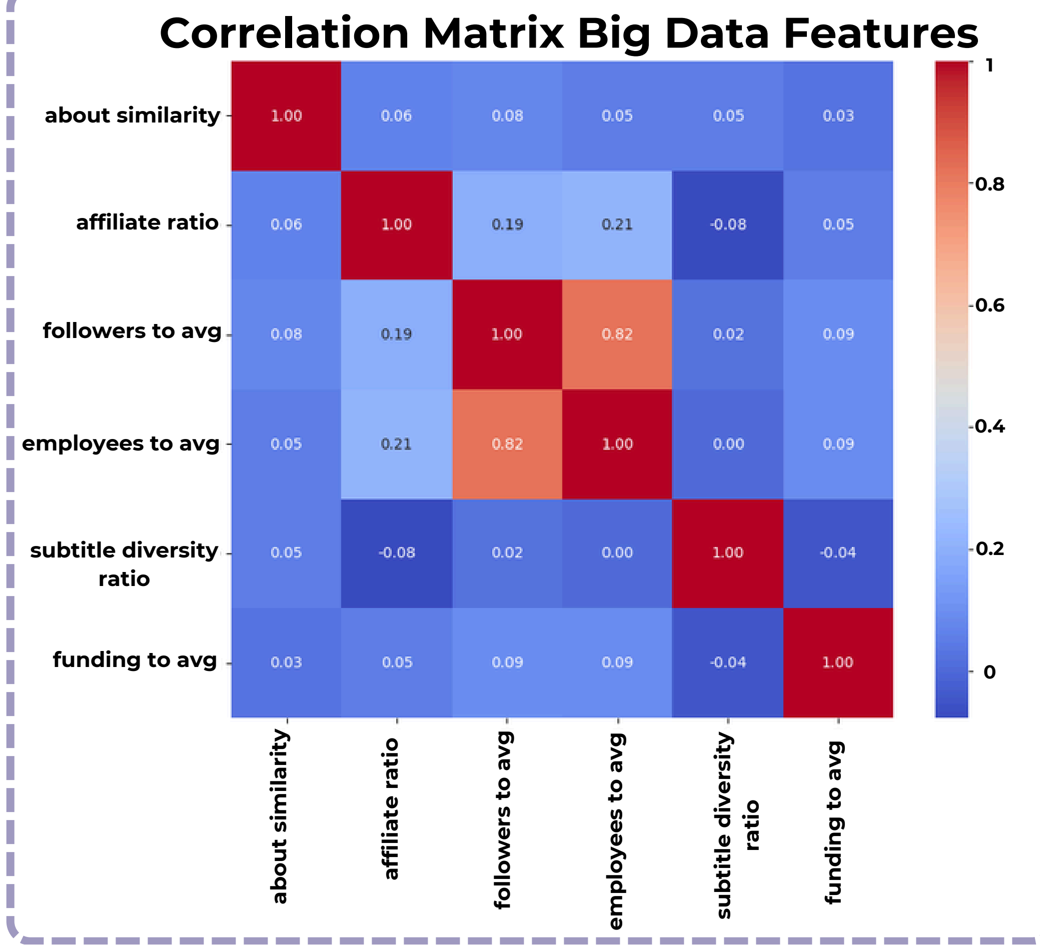
- Clustering
- Employees-to-Average Ratio
- About Similarity

Text to Embedding
SentenceTransformer ('all-MiniLM-L6-v2')

Data Analysis



Word Count Histogram



Correlation Matrix Big Data Features

	about similarity	affiliate ratio	followers to avg	employees to avg	subtitle diversity ratio	funding to avg
about similarity	1.00	0.06	0.08	0.05	0.05	0.03
affiliate ratio	0.06	1.00	0.19	0.21	-0.08	0.05
followers to avg	0.08	0.19	1.00	0.82	0.02	0.09
employees to avg	0.05	0.21	0.82	1.00	0.00	0.09
subtitle diversity ratio	0.05	-0.08	0.02	0.00	1.00	-0.04
funding to avg	0.03	0.05	0.09	0.09	-0.04	1.00

UI

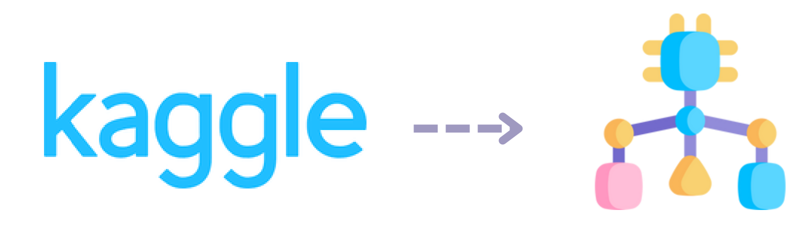
Streamlit

Upon selecting a job listing, users can view information about the job, along with Method 4's model's prediction of whether the job is fraudulent or legitimate.

Conclusions

Big Data feature engineering and semi-supervised learning improve job fraud classification, improving transparency in real-world scenarios.

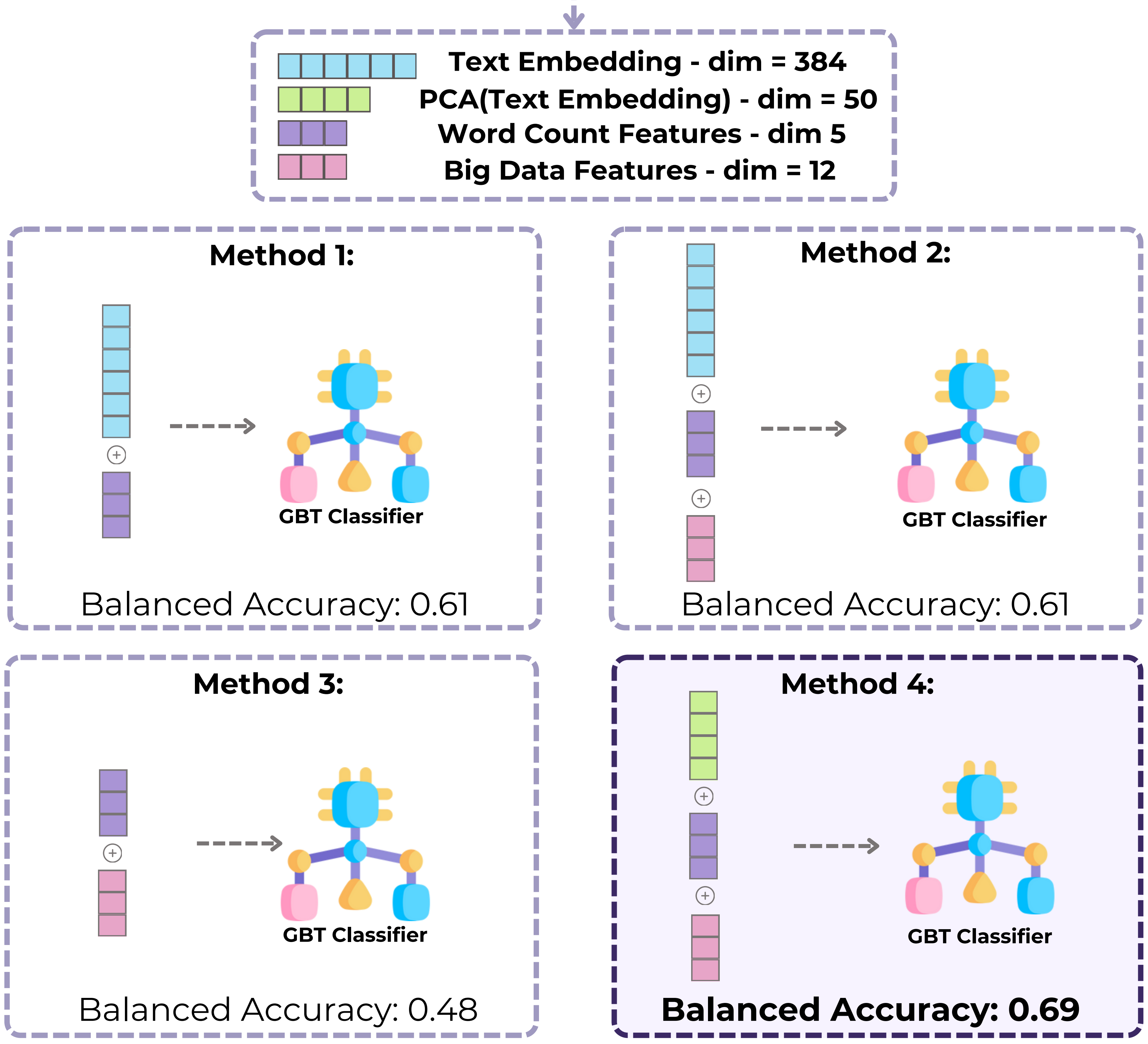
Modeling



Training

Model	F1	Balanced Accuracy
Random forest	0.52	0.67
GBT Classifier	**0.6**	**0.76**
Logistic Regression	0.26	0.57
FF Network	0	0.5

Apply High-Confidence Pseudo-labeling to Scraped Data



- **Text Embedding - dim = 384**
- **PCA(Text Embedding) - dim = 50**
- **Word Count Features - dim 5**
- **Big Data Features - dim = 12**

Method 1:
GBT Classifier
Balanced Accuracy: 0.61

Method 2:
GBT Classifier
Balanced Accuracy: 0.61

Method 3:
GBT Classifier
Balanced Accuracy: 0.48

Method 4:
GBT Classifier
Balanced Accuracy: 0.69